## 8.4 Probabilistic Retrieval Model

- an alternative model for query optimization
- two main parameters – P(REL), probability of relevance and P(NREL), probability of non-relevance of a document.
- the probability that a document d is relevant is given by

$$P(REL \& d) = P(REL) \times P(d / REL)$$

$$= P(d) \times P(REL / d)$$

$$P(REL / d) = P(REL) \times P(d / REL) \ / \ P(d)$$

- given two classes – relevant and nonrelevant, and their probabilities; there are two classes of user judgments and two actions provided by the retrieval system

a. REL (relevant) and NREL (nonrelvant).

b. $\alpha_1$ (retrieve actions) and $\alpha_2$(not retrieve action).

Optimal Decision is

$$P(REL/d) \underset{\alpha_2}{\overset{\alpha_1}{\gtrless}} P(NREL/d)$$

Retrieve *d* if

$$\left(\frac{P(d/REL)}{P(d/NREL)}\right)\left(\frac{P(REL)}{P(NREL)}\right) > 1$$

Taking log,

$$\log\left(\frac{P(d/REL)}{P(d/NREL)}\right) + \log\left(\frac{P(REL)}{P(NREL)}\right) > 0$$

$$P_{\alpha_1 / d}(Error) = P(NREL / d)$$

$$P_{\alpha_2 / d}(Error) = P(REL / d)$$

Average $P(Error)$

$$\sum_{\substack{d \in D \\ \alpha_1 / d}} P(NREL / d) \bullet P(d) + \sum_{\substack{d \in D \\ \alpha_2 / d}} P(REL / d) \bullet P(d)$$

<u>PROBLEM</u>

Find a decision rule and associated q such that Average *P(Error)* is minimized.

# I. Binary "Independence" Model

- documents are represented as

$$d = (w_1, w_2, \ldots, w_t)$$

where each $w_i = \begin{cases} 1 & \text{if term i appears in d} \\ 0 & \text{otherwise} \end{cases}$

- the value of $w_i$ conforms to Bernouli distribution, given by

$$\left.\begin{array}{l} P(w_i = 1 / REL) = p_i \\ P(w_i = 0 / REL) = 1 - p_i \end{array}\right\rangle P(w_i / REL) = p_i^{w_i}(1 - p_i)^{1 - w_i}$$

$$\left.\begin{array}{l} P(w_i = 1 / NREL) = q_i \\ P(w_i = 0 / NREL) = 1 - q_i \end{array}\right\rangle P(w_i / NREL) = q_i^{w_i}(1 - q_i)^{1 - w_i}$$

where $p_i$ and $q_i$ are the probabilities of term $t_i$ in a relevant document and a nonrelevant document respectively.

- because of term independence

$$P(d \,/\, REL) = \prod_{i=1}^{t} p_i^{w_i} (1 - p_i)^{1 - w_i}$$

$$P(d \,/\, NREL) = \prod_{i=1}^{t} q_i^{w_i} (1 - q_i)^{1 - w_i}$$

therefore

$$log\left(\frac{P(d / REL)}{P(d / NREL)}\right) = log\left(\frac{\prod_{i=1}^{t} p_i^{w_i}(1-p_i)^{1-w_i}}{\prod_{i=1}^{t} q_i^{w_i}(1-q_i)^{1-w_i}}\right)$$

$$= \sum_{i=1}^{t} w_i \, log\left(\frac{p_i}{q_i}\right) + \sum_{i=1}^{t}(1-w_i) log\left(\frac{1-p_i}{1-q_i}\right)$$

$$= \sum_{i=1}^{t} w_i \log\left(\frac{p_i}{q_i}\right) - \sum_{i=1}^{t} w_i \log\left(\frac{1-p_i}{1-q_i}\right) + \sum_{i=1}^{t} \log\left(\frac{1-p_i}{1-q_i}\right)$$

$$= \sum_{i=1}^{t} w_i \log\left(\frac{p_i}{q_i} * \frac{(1-q_i)}{(1-p_i)}\right) + \sum_{i=1}^{t} \log\left(\frac{1-p_i}{1-q_i}\right)$$

in the final equation above, factor # 1 alone affects the ranking of the document, factor # 2 is used as a cut-off.

- if $X = (x_1, x_2, \ldots, x_n)$ where each

$$x_i = \log\left(\frac{p_i(1-q_i)}{q_i(1-p_i)}\right)$$

then the rule is to retrieve *d* if

$$d \bullet X^T + c > 0$$

- similar mathematical derivation can be obtained for other distributions like Poisson and Normal distributions.

# Estimating Term Relevance Weights

TABLE 1.

| | $t_1$ | $t_2$ | $t_3$ | $t_4$ | |
|---|---|---|---|---|---|
| $d_1$ | 0 | 1 | 1 | 0 | R |
| $d_2$ | 1 | 0 | 0 | 1 | |
| $d_5$ | 1 | 1 | 0 | 0 | R |
| $d_{10}$ | 1 | 0 | 0 | 1 | |
| $d_{11}$ | 1 | 1 | 1 | 0 | R |

TABLE 1 is used as training data.

$t_1$

| | | REL | NREL |
|---|---|---|---|
| $w_1$= | 1 | 2 | 2 |
| | 0 | 1 | 0 |

$t_2$

| | | REL | NREL |
|---|---|---|---|
| $w_2$= | 1 | 3 | 0 |
| | 0 | 0 | 2 |

$t_3$

| | | REL | NREL |
|---|---|---|---|
| $w_3$= | 1 | 2 | 0 |
| | 0 | 1 | 2 |

$t_4$

| | | REL | NREL |
|---|---|---|---|
| $w_4$= | 1 | 0 | 2 |
| | 0 | 3 | 0 |

Jeffrey's prior

$t_1$

| | | REL | NREL |
|---|---|---|---|
| $w_1$= | 1 | 2.5 | 2 .5 |
| | 0 | 1.5 | .5 |

$x_1 = \log 1/3$

$t_2$

| | | REL | NREL |
|---|---|---|---|
| $w_2$= | 1 | 3.5 | .5 |
| | 0 | .5 | 2.5 |

$x_2 = \log 35$

$t_3$

| | | REL | NREL |
|---|---|---|---|
| $w_3$= | 1 | 2.5 | .5 |
| | 0 | 1.5 | 2.5 |

$x_3 = \log 25/3$

$t_4$

| | | REL | NREL |
|---|---|---|---|
| $w_4$= | 1 | .5 | 2.5 |
| | 0 | 3.5 | .5 |

$x_4 = \log 1/35$